Cisco UCS C845A M8 Rack Server



Overview

The Cisco UCS® C845A M8 Rack Server is Cisco's flagship NVIDIA RTX PRO server platform, purpose-built for enterprise AI inferencing, agentic AI, and the NVIDIA AI Data Platform. It is a highly scalable, flexible, and customizable two- to eight-GPU system based on the NVIDIA MGX reference design for accelerated computing. Together with NVIDIA RTX PRO Blackwell Server Edition GPUs and NVIDIA AI Enterprise software or AMD Instinct MI350P GPUs, it is designed to deliver high performance across multiple AI workloads. As a foundational compute building block for the Cisco Secure AI Factory with NVIDIA, the C845A M8 helps enterprises move from AI pilot to production with a secure, simple, and scalable infrastructure—accelerating time to first token and lowering the cost per token from the core to the edge.



Figure 1. Cisco UCS C845A M8 Rack Server

© 2026 Cisco and/or its affiliates. All rights reserved.

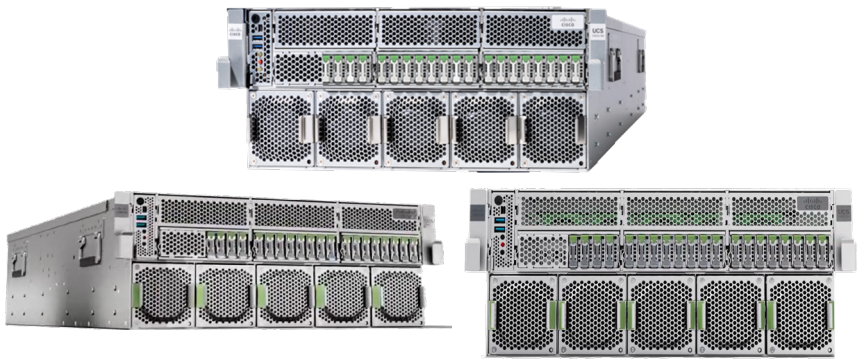


Figure 2. Cisco UCS C845A M8 Rack Server

The versatility of the C845A M8 makes it ideal for a variety of use cases, including:

- **GenAI inferencing, fine-tuning, and RAG:** provides a foundation of knowledge and language patterns, enabling faster adaptation to specific tasks and domains using significantly less data compared to traditional models. With up to 96 GB of GPU memory per NVIDIA RTX PRO 6000 Blackwell GPU or 144 GB of GPU memory per AMD Instinct MI350P GPU and up to 8 TB of system memory, the C845A M8 keeps larger models, longer contexts, and more concurrent users resident in memory, driving higher throughput and lower cost per token for production inferencing.
- **Agentic and physical AI:** powers autonomous agents, smart factories, robotics, and digital twins with NVIDIA RTX PRO Blackwell Server Edition GPUs, delivering real-time decision-making at enterprise scale.
- **NVIDIA AI Data Platform foundation:** serves as the accelerated compute tier for the NVIDIA AI Data Platform, pairing NVIDIA-certified storage partners with NVIDIA AI Enterprise software so enterprises can turn proprietary data into production-grade reasoning, retrieval, and agentic workflows.
- **Core-to-edge distributed inferencing:** deploys low-latency inferencing close to data and users with consistent policy, security, and observability through Cisco Intersight® and the Cisco Secure AI Factory with NVIDIA.
- **High-Performance Computing (HPC):** supports complex computations needed for simulations and large-scale data processing.
- **Data analytics and visualization:** utilizes advanced analytics tools to extract insights from vast datasets, facilitating data-driven decision-making.
- **Design and simulation:** supports 3D content creation and photorealistic simulation workloads such as digital twins, multi-user design collaboration, and Extended Reality (XR).
- **Language processing:** enables servers to understand and interpret human language using complex algorithms, allowing for text analysis, sentiment analysis, machine translation, and human-like text generation.
- **Conversational AI:** utilizes technologies such as natural language processing and machine learning to comprehend human language, discern intent, maintain context within a conversation, and generate human-like responses.
- **Graphics and rendering:** processes large amounts of data swiftly to generate complex visual images, making rendering faster than relying on CPUs.
- **Virtual desktop:** harnesses the power of accelerated computing to boost AI-enhanced virtualized workloads and deliver high-performance workstation instances to remote users with Virtual Desktop Infrastructure. (VDI). Ideal for remote workloads with CAD, video editing, 3D modeling, and AI use cases.

© 2026 Cisco and/or its affiliates. All rights reserved.

Benefits

Optimized for AI use cases

Built on NVIDIA MGX modular reference design, the Cisco UCS C845A M8 Rack Server provides the accelerated computing power from NVIDIA RTX PRO Blackwell and NVIDIA AI Enterprise software or AMD Instinct MI350P GPUs necessary to handle the most challenging AI workloads. The platform is engineered for the AI era, optimized for high-throughput, memory-bound inferencing of LLMs, agentic AI, and multimodal models, with NVIDIA RTX PRO 6000 (96GB), RTX PRO 6000D (84GB), and RTX PRO 4500 (32GB) Blackwell Server Edition GPUs or AMD Instinct MI350P (144GB) GPUs to right-size every deployment.

Adaptable design

With flexible options for two, four, six or eight NVIDIA RTX PRO 4500 (32GB), RTX PRO 6000D (84GB), and/or 6000 (96GB) Blackwell GPUs or AMD Instinct MI350P (144GB) GPUs, you can begin with a smaller setup and expand as your needs grow. Up to 8 TB of DDR5-6400 system memory and Gen5 E1.S NVMe storage (up to 15.3 TB per drive) keep large models, vector indexes, and KV caches close to the GPU, sustaining high tokens-per-second under real workloads. The modularity of the NVIDIA MGX design supports a wide variety of use cases.

Secure, simple, and scalable AI infrastructure

As a building block of the Cisco Secure AI Factory with NVIDIA, the C845A M8 inherits distributed security fused into every layer, protecting models, agents, workloads, and infrastructure with Cisco AI Defense, Hybrid Mesh Firewall, Isovalent®, and Hypershield™. NVIDIA BlueField-3 DPUs with crypto acceleration offload security policy enforcement from CPU and GPU, preserving cycles for AI processing while keeping the entire stack hardened and observable through Splunk® Enterprise Security and Splunk Observability Cloud.

Consistent management

Manage your AI infrastructure seamlessly with Cisco Intersight, an operations platform that helps IT teams see, control, and automate their Cisco UCS, converged, and hyperconverged infrastructure throughout its lifecycle—wherever it is—from one place.



© 2026 Cisco and/or its affiliates. All rights reserved.

What it offers

The Cisco UCS C845A M8 Rack Server, built on the NVIDIA MGX reference design for accelerating computing, brings AI capabilities to mainstream enterprise PCIe servers. Its adaptable configuration addresses a variety of data-center workloads, from demanding Generative AI use cases to more mainstream graphics-accelerated VDI solutions.

Configurations

- Two 5th Gen AMD EPYC CPUs in a 4RU form factor, with up to 8 TB of DDR5-6400 system memory using 32GB to 256GB DIMMs.
- 2, 4, 6 or 8x NVIDIA RTX PRO 4500 (32 GB), RTX PRO 6000D (84 GB), and 6000 Blackwell (96 GB), H200 NVL/H100 NVL/L40S, AMD Instinct MI350P and AMD MI210 GPUs.
- 5x PCIe x16 FHHL slots and 8 x PCIe x16 GPU slots.
- 4x single-port 400G NVIDIA ConnectX-7 Smart NICs or NVIDIA BlueField-3 DPUs with crypto-enabled options or AMD Pollara 400G AI NIC for in-line data protection to scale out. One dual port 200G NIC or DPU to scale up for north/south traffic.
- Up to 20x Gen5 E1.S NVMe SSDs (up to 15.3 TB per drive) for high-speed local storage.

Software

Systems equipped with NVIDIA H100 NVL or H200 NVL come with a 5-year license for NVIDIA AI Enterprise, a cloud-native software platform that streamlines development and deployment of production-grade AI solutions, including AI agents, Generative AI, computer vision, speech AI, and more. Easy-to-use microservices optimize model performance with enterprise-grade security, support, and stability, ensuring a smooth transition from prototype to production for enterprises that run their businesses on AI.

The C845A M8 is engineered to align with the NVIDIA AI Data Platform reference design, connecting enterprise data, vector stores, and retrieval pipelines to GPU-accelerated inferencing, so customers can build agentic AI and RAG applications on their own data with the performance, governance, and security required for production. Combined with NVIDIA NIM microservices, NeMo, and Blueprints, the platform compresses time to first token and shortens the path from pilot to enterprise deployment.

Management

The Cisco UCS C845A M8 Rack Server is managed by Cisco Intersight, a cloud-delivered IT operations platform that helps your IT operations team see, control, and automate the Cisco UCS infrastructure throughout its lifecycle—wherever it is—from one place.

By using Intersight, you can operate with consistency and control, strengthen your security posture, and increase energy efficiency to drive innovation and growth.

Learn more

For additional information about the Cisco UCS C845A M8 Rack Server, refer to the data sheet.

For information about our data center solutions for AI visit https://www.cisco.com/site/us/en/solutions/artificial-intelligence/infrastructure/index.html.

© 2026 Cisco and/or its affiliates. All rights reserved. Cisco and the Cisco logo are trademarks or registered trademarks of Cisco and/or its affiliates in the U.S. and other countries. To view a list of Cisco trademarks, go to this URL: www.cisco.com/go/trademarks. Third-party trademarks mentioned are the property of their respective owners. The use of the word partner does not imply a partnership relationship between Cisco and any other company. (1110R)

C45-4892551-04 09/26